MGMTMSA 437 Time Series Analysis and Forecasting

Hotel Demand Forecasting

Jiayi Zhao, Jiachen Yu, Zhi Zhang, Sen Yang

Executive Summary

This report presents an analysis of hotel booking demand using historical data to develop predictive models that can inform operational decisions in the hospitality industry. Given the variability in demand due to seasonality, economic trends, and market events, accurate forecasting is essential for optimizing resource allocation, maximizing occupancy, and improving customer experience. Using a range of time series and machine learning models, we evaluate the performance of each approach, ultimately concluding that a simple linear regression (OLS) model outperforms more complex models. The findings suggest that focusing on key features, such as average daily rate (ADR) and GDP, enables reliable demand forecasts, offering practical insights for hotel management.

Purpose/Motivation

The primary objective of this project is to forecast the number of hotel bookings over time using historical booking data. Demand in the hospitality industry fluctuates significantly due to factors such as seasonality, economic cycles, and special events. Accurate forecasts are crucial for hotels to optimize resource allocation, tailor marketing strategies, and enhance customer experience. This project aims to deliver reliable, actionable insights into booking patterns, enabling hotels to make informed, data-driven decisions that drive operational efficiency and improve revenue management.

This analysis employs advanced time series models to capture the inherent patterns and trends in booking data, allowing for precise short-term and long-term forecasts. Model performance will be rigorously evaluated through test-set validation to ensure robustness, with visualizations highlighting predictive accuracy and key takeaways from the findings. Beyond improving booking prediction accuracy, this project illustrates the practical application of machine learning

in the hospitality industry, demonstrating the potential of data analytics to enhance business operations.

Data Description, Preprocessing, and Exploratory Data Analysis (EDA)

Data Description

The dataset, sourced from Kaggle Hotel Booking Demand Dataset, contains detailed records of hotel bookings, including temporal variables crucial for time series forecasting. Key features include booking count, arrival date, average daily rate (ADR), guest numbers, and contextual data such as GDP and holiday counts. These variables provide insights into demand drivers and allow for a nuanced analysis of booking patterns.

Data Preprocessing

- **Date Conversion**: We combined year, month, and day columns into a unified arrival_date field in datetime format to facilitate time-based aggregation and analysis.
- **Guest Count Calculation**: We calculated the total number of guests per booking by summing the counts of adults, children, and babies, creating a guest_num feature that represents overall occupancy per booking.
- **Aggregation**: To reduce noise and highlight trends, data was aggregated by arrival_date, computing mean values for features like stays_in_weekend_nights, stays_in_week_nights, guest_num, and adr. The total number of bookings per date was calculated as booking_number.
- **Integration of GDP Data**: GDP data was integrated as an economic indicator to provide context about the broader economic environment, given that economic conditions can impact travel demand. Using the Fred API, we retrieved GDP data, matched it with hotel booking data by month and year, and forward-filled missing values to maintain continuity.
- **Weekly Resampling**: The data was resampled to a weekly frequency, smoothing day-to-day fluctuations and providing a more stable basis for forecasting.
- **Holiday Feature Creation**: A holiday_count feature was added to capture the number of U.S. holidays in each week, reflecting potential surges in demand around holiday periods.

```
df
```

	arrival_date	booking_number	stays_in_weekend_nights	stays_in_week_nights	guest_num	adr	GDP	holiday_count
0	2015-07-05	412	1.253678	3.203693	2.029519	91.560474	18401.626	2
1	2015-07-12	495	1.287284	3.218006	2.150637	94.995764	18401.626	0
2	2015-07-19	684	1.097910	2.977407	2.026222	103.977369	18401.626	0
3	2015-07-26	763	1.065040	2.836771	2.108315	103.529133	18401.626	0
4	2015-08-02	582	1.225607	3.217633	2.132673	118.629752	18401.626	0
...	...	...	...	...	...	...	...	...
109	2017-08-06	1146	1.189595	3.101027	2.310544	168.404240	19692.595	0
110	2017-08-13	1091	1.178829	3.014282	2.293373	166.101119	19692.595	0
111	2017-08-20	1241	1.075447	2.892131	2.299406	167.089947	19692.595	0
112	2017-08-27	1070	0.995120	2.744319	2.264675	162.883080	19692.595	0
113	2017-09-03	559	0.711073	3.323672	2.050856	149.273742	19692.595	0

114 rows × 8 columns

Figure 1. Data after Cleaning

Exploratory Data Analysis (EDA)

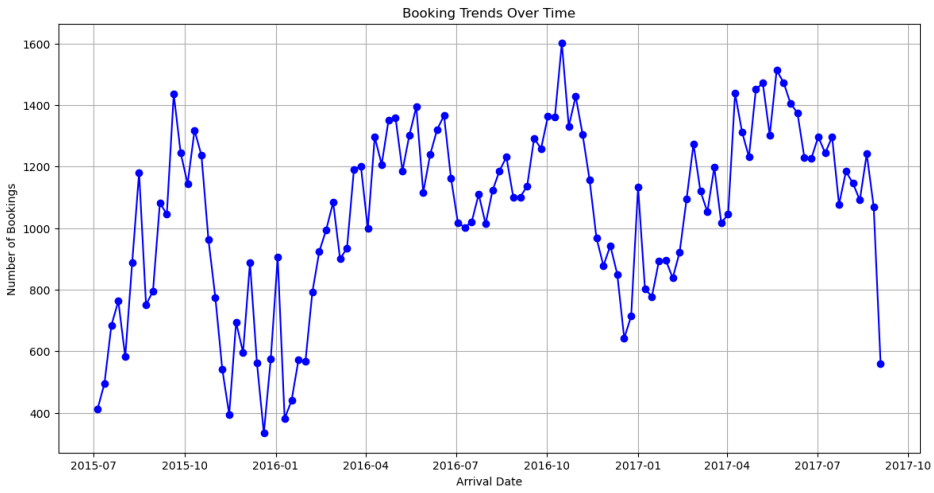

Figure 2. Booking Trends Over Time

Figure 2 depicts booking counts over time, revealing seasonal peaks, especially during high-demand periods such as summer and winter holidays. The observed cyclicality suggests that models capable of capturing seasonality, like ARIMA or seasonal variants of VAR, may be effective. The overall trend shows a relatively stable demand with recurrent spikes, which aligns with typical patterns in the hospitality industry.

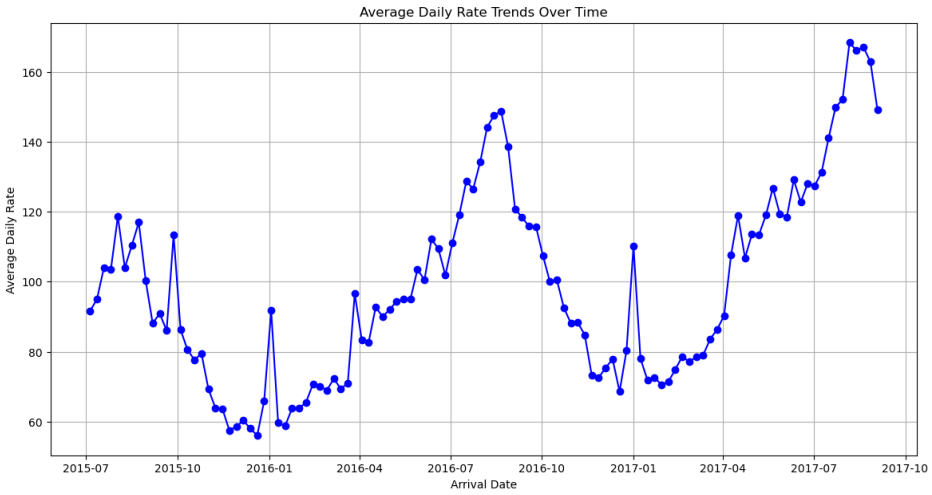

Figure 3. Average Daily Rate (ADR) Trends Over Time

ADR, shown in Figure 3, reflects the average rate per booking over time. The cyclical pattern in ADR aligns with demand peaks, indicating that hotels adjust pricing in response to expected occupancy levels. This observation suggests ADR as a valuable feature for forecasting models, given its alignment with demand trends. High ADR periods, coinciding with peak seasons, further validate the relationship between demand and pricing adjustments in hospitality.

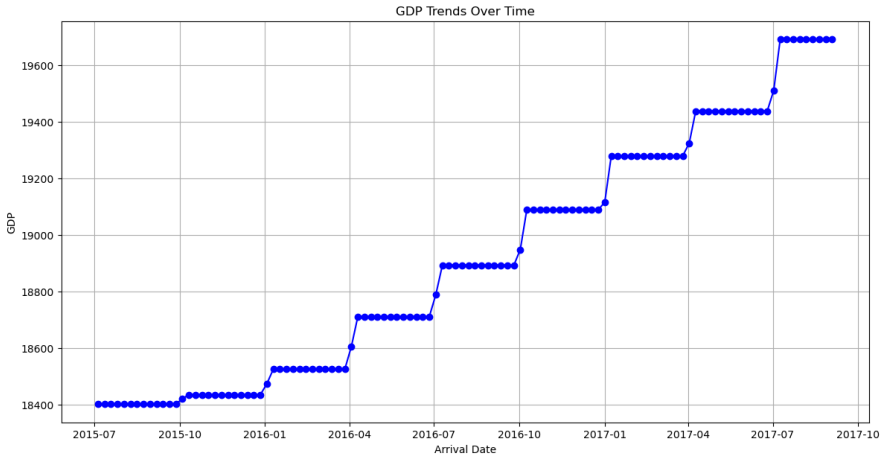

Figure 4. GDP Trends Over Time

GDP, included as an economic indicator, shows a steady increase over the observed period, mirroring economic growth trends. This upward trajectory implies that GDP could have a positive relationship with booking counts, as economic growth typically correlates with increased discretionary spending, including travel. Integrating GDP into multivariate models like VAR allows the model to account for this broader economic context.

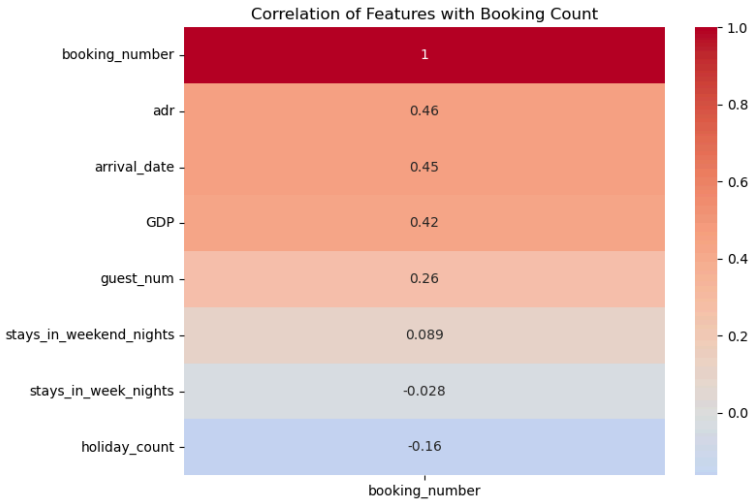

Figure 5. Correlation of Features with Booking Count

The correlation heatmap visualizes the relationships between various features in the dataset and booking numbers. Booking numbers show a strong positive correlation with ADR and arrival date, indicating that higher rates and certain times correlate with increased bookings. The correlation with GDP is also positive, suggesting an economic influence on booking trends. Features like holiday count and weekday stays exhibit weak or negative correlations with booking numbers. This may imply that holiday seasons or longer stays during weekdays have less impact on weekly booking totals.

Train-Test Split

To ensure realistic model evaluation, an 80-20 chronological split was applied, with the first 80% of the data used for training and the final 20% reserved for testing. This split preserves temporal order, allowing us to validate the model on unseen, future data. The training set spans from 2015-07-05 to 2017-04-02, while the test set covers 2017-04-09 to 2017-09-03. This setup ensures that the model is tested on a period that follows the training data, simulating real-world forecasting conditions where future data is not available during model training.

Stationarity Check

Stationarity is a crucial assumption in time series modeling, especially for ARIMA and VAR models. We performed the Augmented Dickey-Fuller (ADF) test on each time series to assess stationarity. For non-stationary series, transformations were applied as follows:

- Differencing: Applied to variables like booking count and ADR to remove trends and achieve a stable mean, which is essential for accurate time series forecasting.
- Growth Calculation: For some features, we calculated the growth rate, providing a stabilized series that captures changes over time without absolute fluctuations.

By ensuring stationarity, we prepared the data for effective use in autoregressive models. The stationarity-adjusted series, as shown in Figure 6, confirm the successful transformation of key variables, with ACF plots reflecting reduced autocorrelation and stable means, which are essential for reliable model training.

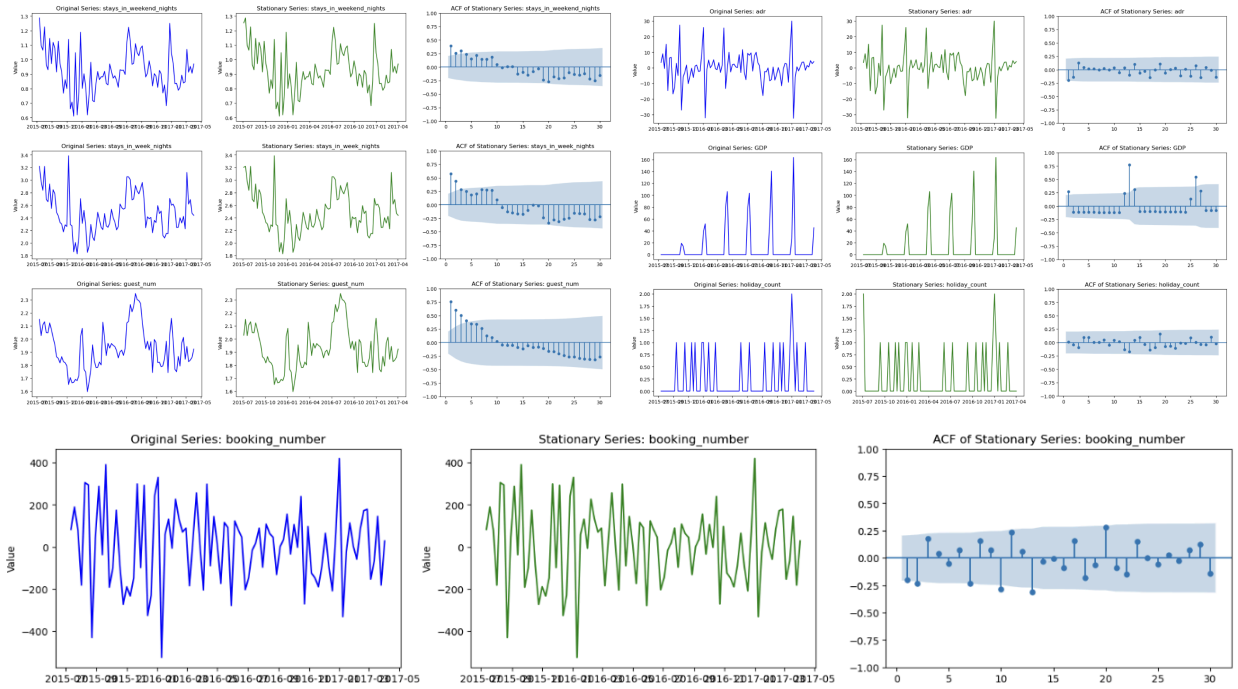

Figure 6. Stationarity Transformation and Autocorrelation Analysis of Time Series Variables

Methodology and Model Selection

In this section, various models are selected to capture the dynamics of the data. Each model is described in terms of its suitability for the dataset and predictive goals.

Ordinary Least Squares (OLS) Model

Serving as a baseline, the OLS model provides a simple linear regression forecast to capture linear relationships between booking numbers and key predictors like ADR, GDP, and holiday counts. Despite its limitations in capturing non-linear and seasonal effects, OLS offers a straightforward approach to gauge linear trends in booking demand. Figure 7 illustrates that while OLS captures general trends, it struggles with seasonal peaks, highlighting its limitations in modeling cyclic behavior. It achieved an R-squared of 0.143, indicating limited explanatory power. However, it performed well in forecasting, with a Mean Squared Logarithmic Error (MSLE) of 2.35 on the test set, the lowest among all models. This low MSLE suggests that despite its simplicity, the OLS model accurately captures proportional variations in booking demand.

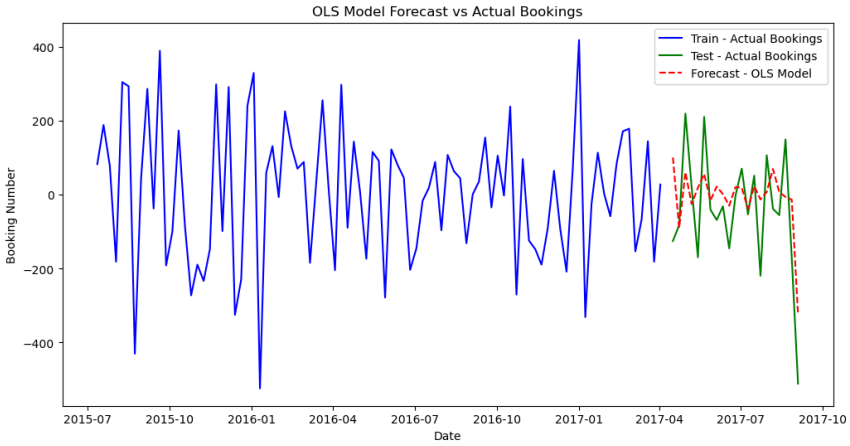

Figure 7. OLS Model Forecast vs Actual Bookings

VAR Models

The Vector Autoregression (VAR) models allow us to capture the relationships among multiple time series, such as bookings, ADR, and GDP. Two variants were tested:

- VAR Level Model: This model considers the raw levels of each variable, assuming that they influence each other directly over time. After selecting the optimal lag based on AIC criteria, the model forecasted future booking numbers. However, the VAR Level model had a relatively high MSLE of 12.79 on the test set, suggesting that while it captures interdependencies, it may not be as effective in predicting the absolute values of bookings. As we can see from the Figure 8, this is a bad performance. The underperformance of the VAR Level model may be due to its complexity, which might have led to overfitting given the limited amount of data points, especially in a multivariate setting.

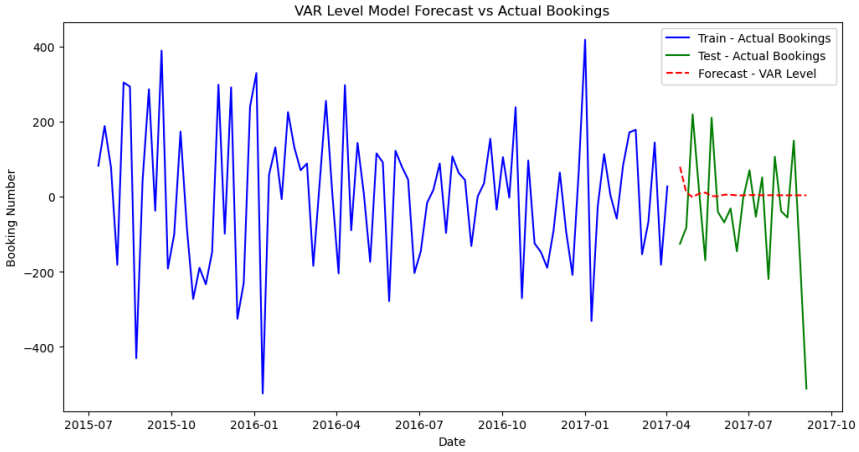

Figure 8. VAR Level Model Forecast vs Actual Bookings

- VAR Growth Model: This variant relies on the growth rates (differenced data), making it better suited for capturing relative changes rather than absolute levels. The MSLE for the VAR Growth model was slightly lower than the VAR Level model, at 12.07, but it was still

significantly higher than the OLS model. This result indicates that while differencing stabilizes the series, it may lead to underestimation of peak demand periods due to its emphasis on smooth changes. This suggests that growth-based modeling failed to capture the stable, linear trends adequately, potentially due to an overemphasis on short-term variability rather than the overall booking trends.

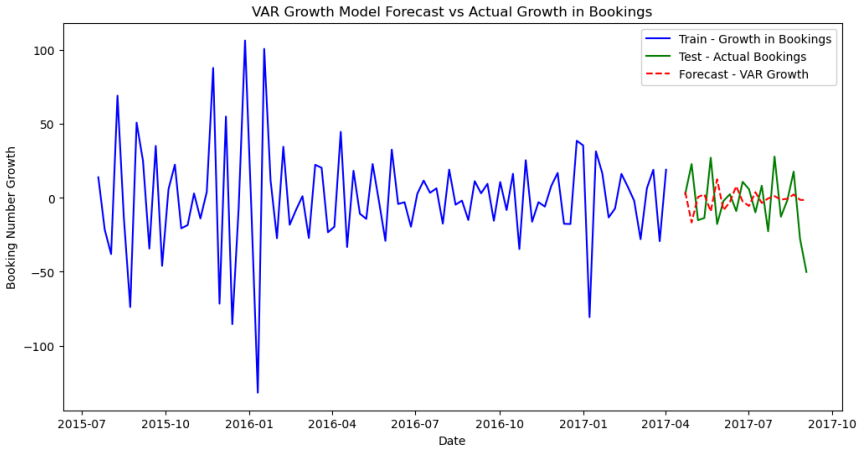

Figure 9. VAR Growth Model Forecast vs Actual Growth in Bookings

ARIMA Model

The ARIMA model was optimized using auto_arima, which automatically selected parameters to minimize AIC. The best-fitting ARIMA model included seasonal terms to capture the cyclic nature of hotel bookings. However, surprisingly, the ARIMA model achieved an MSLE of 21.18, the worst among all models. Despite ARIMA's strengths in handling autocorrelations and temporal dependencies, its underperformance may stem from the fact that it overly relied on past values without effectively incorporating the external predictors (e.g., ADR and GDP). Additionally, ARIMA may have struggled with the multivariate nature of the dataset, as it was only applied to the primary target variable (booking numbers) without capturing the full interplay between all influencing factors.

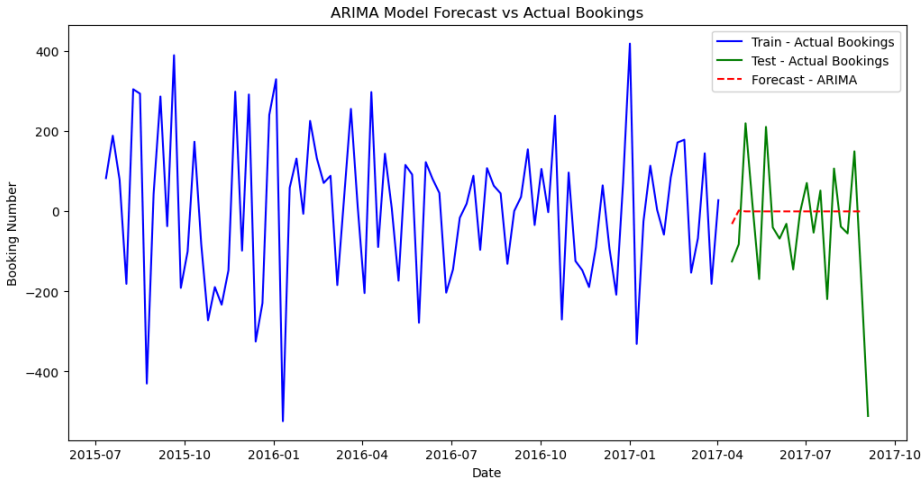

Figure 10. ARIMA Model Forecast vs Actual Bookings

Results and Model Comparison

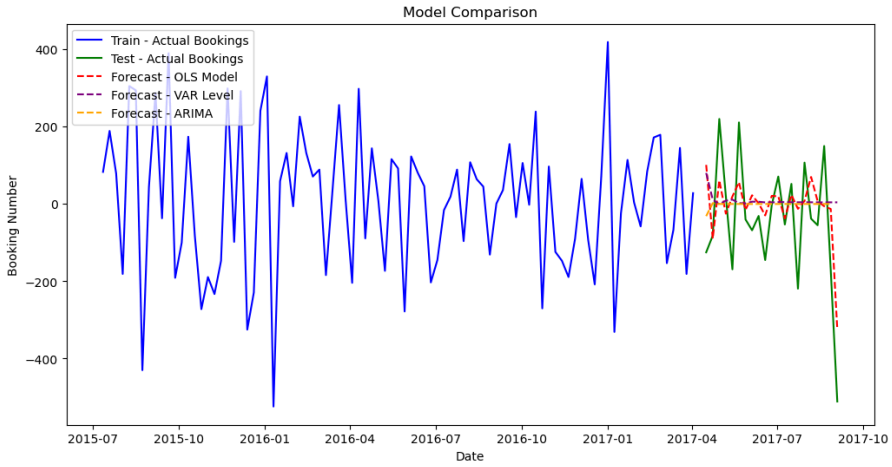

Figure 11. Model Comparison

Figure 11 provides a comprehensive forecast comparison for each model against actual bookings. Among the models, the OLS model, despite its simplicity, yielded the lowest MSLE score of 2.35, indicating better alignment with test data. While VAR models offered insights into multivariate dynamics, their MSLE values (12.79 for VAR Level and 12.07 for VAR Growth) were considerably higher, demonstrating limited accuracy in forecasting absolute booking numbers. The ARIMA model, with an MSLE of 21.18, had the weakest performance due to its inability to account for multivariate relationships.

The superior performance of the OLS model highlights the strength of simpler models when the relationships between variables are primarily linear and not dominated by complex temporal dynamics. This result underscores that, in certain cases, simpler models can achieve better generalization by avoiding overfitting, especially in datasets with relatively stable linear

relationships. The findings suggest that while advanced time series models like VAR and ARIMA are powerful, they may not always be the optimal choice for datasets where linear trends and direct feature relationships dominate. The success of the OLS model demonstrates that in forecasting tasks for hotel bookings with similar features, linear regression may be a practical and effective approach, enabling hotels to make data-driven decisions with high predictive accuracy and reduced computational complexity.

Conclusions and Implications

Our analysis concludes that the OLS model, with the lowest MSLE, is the most practical choice for forecasting short-term hotel bookings in this dataset. This finding underscores the potential of linear models when combined with thoughtful feature selection, such as ADR and GDP. The implications for hotel management are significant, as a straightforward OLS model allows for easy implementation and reliable predictions, aiding in strategic staffing, inventory management, and pricing adjustments. Future work could incorporate additional external data, such as events or weather patterns, to enhance forecast accuracy over extended horizons.

This study underscores that effective forecasting does not always require advanced models. By focusing on relevant features and applying a straightforward model, we achieved a forecast that is easy to interpret and implement, providing reliable support for operational decisions in the hotel industry. The OLS model's performance reaffirms that in data-driven forecasting, simplicity and feature relevance can be as powerful as model complexity.